Cognitive Computation (2024) 16:2264–2278
https://doi.org/10.1007/s12559-024-10306-z

RESEARCH

Detection of Cardiovascular Diseases Using Data Mining Approaches: Application of an Ensemble-Based Model

Mojdeh Nazari[1] · Hassan Emami[1] · Reza Rabiei[1] · Azamossadat Hosseini[1] · Shahabedin Rahmatizadeh[1]

Received: 30 September 2023 / Accepted: 12 May 2024 / Published online: 30 May 2024

Abstract

Cardiovascular diseases are the leading contributor of mortality worldwide. Accurate cardiovascular disease prediction is crucial, and the application of machine learning and data mining techniques could facilitate decision-making and improve predictive capabilities. This study aimed to present a model for accurate prediction of cardiovascular diseases and identifying key contributing factors with the greatest impact. The Cleveland dataset besides the locally collected dataset, called the Noor dataset, was used in this study. Accordingly, various data mining techniques besides four ensemble learning-based models were implemented on both datasets. Moreover, a novel model for combining individual classifiers in ensemble learning, wherein weights were assigned to each classifier (using a genetic algorithm), was developed. The predictive strength of each feature was also investigated to ensure the generalizability of the outcomes. The ultimate ensemble-based model achieved a precision rate of 88.05% and 90.12% on the Cleveland and Noor datasets, respectively, demonstrating its reliability and suitability for future research in predicting the likelihood of cardiovascular diseases. Not only the proposed model introduces an innovative approach for specifying cardiovascular diseases by unraveling the intricate relationships between various biological variables but also facilitates early detection of cardiovascular diseases.

Keywords Cardiovascular diseases · Heart · Data mining · Classification techniques · Ensemble learning

Introduction

Nowadays, with the emergence of large amounts of data in various fields, collecting data and extracting useful information from data are considered challenging, and data mining is regarded as a solution for the effective utilization of large data sources. Data mining approaches could bring this opportunity for time-restricted clinicians to identify patterns, trends, potential outbreaks, and ultimately effective prevention and control strategies. However, the extensive volume of data besides the resulting uncertainties presents challenges in obtaining dependable results. Data mining has the potential to overcome existing limitations and can help to extract crucial details that can aid in the early detection of diseases [1, 2].

Data mining can assist disease prediction by applying contributing factors extracted from data sources including patients' medical history. By increasing the number of attributes involved in the diagnosis of diseases, it becomes increasingly challenging even for skilled medical practitioners to diagnose and predict outcomes. Consequently, in recent decades, computer-based decision support tools have been widely used to aid physicians in reducing medical errors resulting from fatigue, lack of adequate experience, and the burden of workloads. Using data mining, physicians can analyze medical data more efficiently, with greater precision and detail, within a shorter timeframe [3, 4].

The World Health Organization (WHO) states that cardiovascular diseases are the primary reason for fatalities worldwide. Each year, an estimated 17.9 million people die from cardiovascular diseases, accounting for about 31% of all global deaths. Recent data from the American Heart Association reveals that coronary heart disease remains the leading cause of death in the USA in 2022. On the other hand, cardiovascular diseases not only impact mortality rates

✉ Hassan Emami
haemami@sbmu.ac.ir

✉ Reza Rabiei
r.rabiei@sbmu.ac.ir

[1] Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

but also contribute to significant morbidity, disability, and reduced quality of life [5]. Fortunately, the prevention of cardiovascular diseases is possible by avoiding detrimental factors, such as unhealthy diets, sedentary lifestyles leading to overweight and obesity, the harmful use of tobacco, and excessive alcohol consumption. As a result, individuals with elevated risks of cardiovascular disease, attributable to various factors like high cholesterol levels, chest pain, hypertension, and diabetes, require early diagnosis mechanisms to manage their general health conditions and avert unexpected heart failure.

The utilization of data mining introduced a fresh perspective on predicting cardiovascular diseases. Accordingly, different data mining techniques were employed to identify and extract valuable information from clinical datasets with minimal user input and effort. In recent years, researchers have explored diverse approaches to implement data mining in healthcare to obtain precise predictions of cardiovascular diseases [6–8]. Dwivedi [9] implemented six different machine-learning classification methods on the Statlog heart disease dataset. Bhatt et al. [10] employed two classification techniques, namely J48 on the Hungarian dataset and Naïve Bayes on the echocardiogram dataset. Sarangam [11] applied and compared four different classification methods on the Cleveland dataset for the prediction of heart disease. Based on a review of 25 studies that leveraged the Cleveland dataset as the baseline [12], various classification methods were implemented, examined, and compared to determine the best-performing method. Briefly, using data mining for cardiovascular disease prediction leads to early detection, improved accuracy, personalized medicine, enhanced decision-making, and effective public health planning, all of which contribute to better patient outcomes and the overall management of cardiovascular health at both individual and population levels [6, 13, 14].

Although numerous researches have been conducted in this area [15–19], there is still a lack of a precise predictive model that can effectively recognize all the critical attributes of cardiovascular diseases. Considering the rising number of individuals afflicted with cardiovascular diseases and the potential of data mining methods to predict these conditions using available data, we decided to utilize the Crisp data mining (Crisp-DM) methodology [20] to create a decision support system framework. In other words, while the detection of cardiovascular diseases requires different tests, this particular model aims to assist physicians in predicting cardiovascular diseases based on each patient's general characteristics.

In summary, this study aimed to determine the most important features and the most suitable data mining techniques for predicting cardiovascular disease besides investigating the efficiency of ensemble learning in increasing the overall performance. Accordingly, various experiments were carried out to identify these features and techniques. To this end, two different datasets were used in our implementation. The first dataset was obtained from the UCI machine learning repository, namely the Cleveland dataset,[1] due to its widespread usage among machine learning researchers and its comprehensive record completeness [21]. The second one was a local dataset collected from the medical information of patients, who visited Noor Heart Center, which is the largest specialized center for heart diseases in the north of Iran, where more than 200 people are served on a daily basis for checkups. The collected dataset features are identical to the Cleveland dataset. The contribution of this paper can be summarized as follows:

- Locally collected dataset besides the Cleveland dataset was used in our experiments to determine the most important features and the most suitable data mining method for predicting cardiovascular disease.
- Crisp-DM methodology was used to make a decision support system framework aiming to increase the success rate of data mining methods.
- Various data mining methods were first implemented on both datasets, and thereafter, voting algorithm, a representative of ensemble learning, is used to combine individual classification methods to classify new instances.
- A weighted majority vote based on the genetic algorithm was utilized to increase the voting algorithm's performance.
- Based on the empirical results, a reliable, accurate, and thorough framework for cardiovascular disease prediction is proposed which not only could play a significant role in resource management and utilization but also could be used by cardiologists as an invaluable and convenient instrument to classify newly diagnosed patients.

The remainder of this paper is organized as follows: The employed methodology including dataset description proposed clinical decision support systems, and its details are mentioned in the "Methodology" section. The "Experiment and Results" section includes the results of the experiments. Discussion and conclusion are respectively provided in the "Discussion" and "Conclusion" sections.

Methodology

Datasets

As previously mentioned, two various datasets were used in our experiments. The first one was the Cleveland dataset collected from the UCI machine learning repository.

[1] https://archive.ics.uci.edu/dataset/45/heart+disease

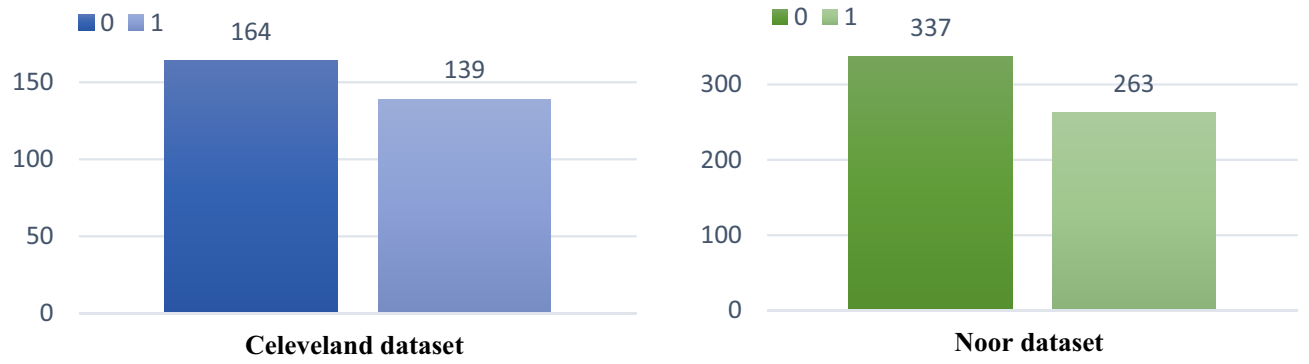


Fig. 1 Distribution of "*Num*" attribute on Cleveland and Noor datasets

This dataset has been extensively used by machine learning experts and contains exceptionally comprehensive records. To provide a more robust basis for data analysis, we decided to collect a local dataset with the same attributes as Cleveland. Accordingly, we collected the medical data of patients who visited Noor Heart Center from April to June 2023. The second dataset called the "Noor dataset" is freely available for academic purposes upon request.

Both datasets contained 14 attributes while 13 of them were used as heart disease prediction features and one was the nominator of output or predicted attribute for the absence or presence of heart disease in a patient. There existed an attribute called "*Num*" in the Cleveland dataset which denoted the heart disease diagnosis in patients on a range of scales, spanning from 0 to 4. In this context, a value of 0 signified the absence of heart disease, while values ranging from 1 to 4 indicated the presence of heart disease (higher values corresponded to greater severity of the condition). To simplify the predicted attribute for the absence or presence of heart disease in the Cleveland dataset, a transformation was applied to convert the multi-class values (0 for absence and 1, 2, 3, and 4 for presence) into binary values which involved converting all diagnosis values from 2 to 4 into 1. As a result, the Cleveland dataset only consisted of the values 0 and 1, where 0 shows the absence and 1 shows the presence of heart disease. Accordingly, to collect the Noor dataset, only 0 and 1 were considered the value of the "*Num*" attribute. Moreover, Cleveland and Noor datasets respectively included 303 and 600 samples. The distribution of the "*Num*" attribute among all records in both datasets is provided in Fig. 1. The details of the attributes and their possible values are described as follows. Notably, all records with missing values were eliminated from both datasets. While Cleveland and Noor datasets respectively had 6 and 11 missing values, their number of records was reduced to 297 and 589. The distribution of continuous features and histogram of discrete features of both datasets are respectively provided in Table 1 and Figs. 2 and 3.

1. Age: This feature indicates the age of the patient in an admitted year.
2. Sex: This binary feature represents whether the patient is male (1) or female (0).
3. Cp: This feature shows the type of chest pain which can have typical angina (1), atypical angina (2), non-angina pain (3), and asymptomatic (4) values.
4. Trestbps: This numeric feature indicates the resting blood pressure on admission to the hospital (mm/Hg).
5. Chol: This numeric feature shows serum cholesterol (mg/dl).
6. Fbs: This binary feature shows fasting blood sugar > 120 (mg/dl) that can have true (1) and false (0) values.
7. Restecg: This feature shows the resting electrocardiographic results with three values of normal (1), abnormal (1), and probable (2).

Table 1 Distribution of numerical features in both datasets (Cleveland Noor datasets)

Feature name	Min		Max		Average		Standard deviation	
	Cleveland	Noor	Cleveland	Noor	Cleveland	Noor	Cleveland	Noor
Age	29	28	77	85	54.54	53.20	9.05	9.63
Trestpbs	94	99	200	188	131.69	124.23	17.76	15.48
Chol	126	124	564	543	247.35	253.54	51.99	49.85
Thalach	71	73	202	193	149.59	150.16	22.94	19.28
Oldpeak	0	0	6.2	5	1.05	0.82	1.16	1.01

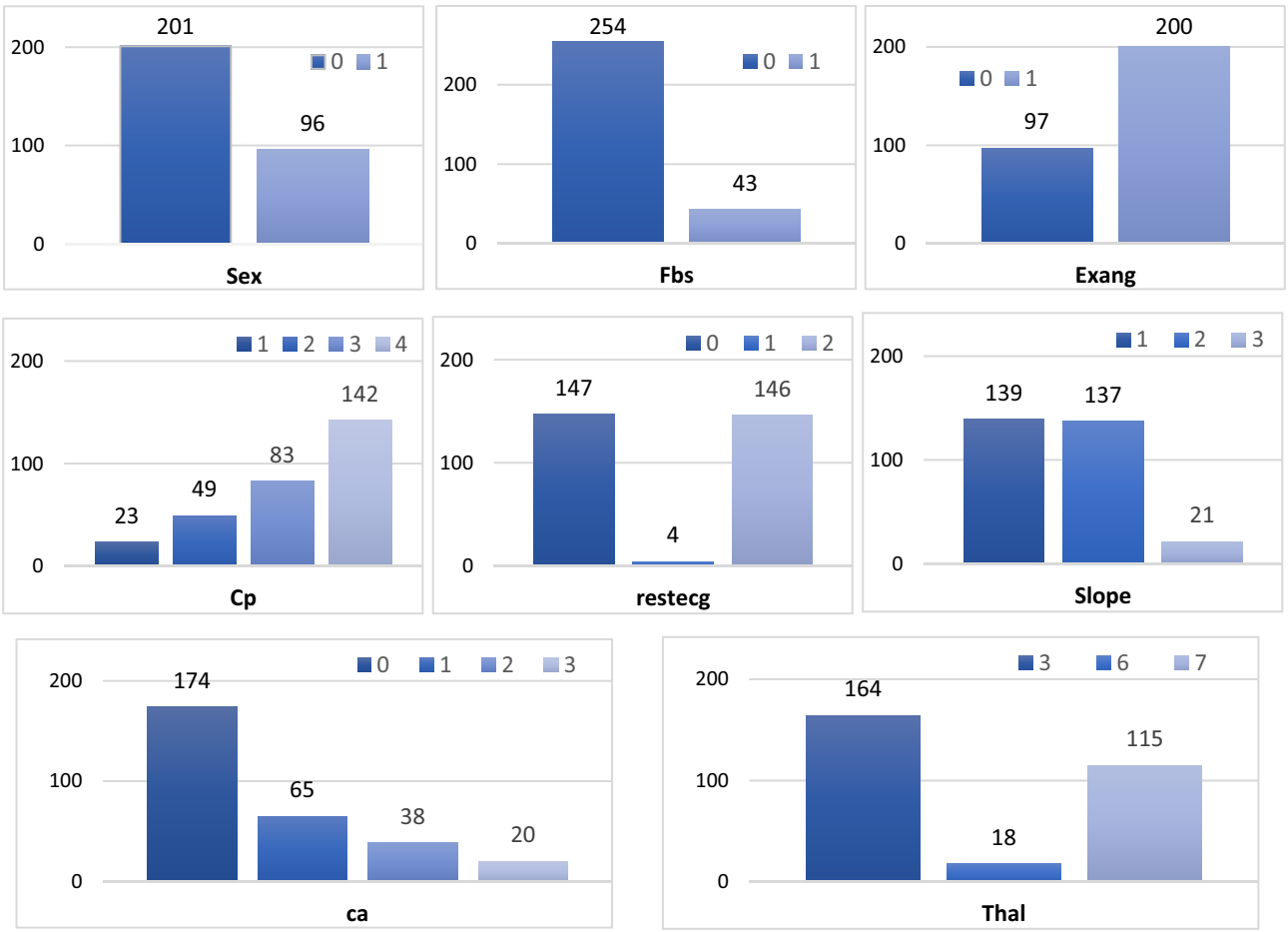


Fig. 2 Histogram of nominal features distribution in the Cleveland dataset

8. Thalach: This numeric feature indicates the maximum heart rate.
9. Exang: This binary feature shows exercise-induced angina with values of yes (1) and no (0).
10. Oldpeak: This numeric feature shows ST depression induced by exercise relative to rest.
11. Slope: This feature shows the slope of the peak exercise ST segment with upsloping (1), flat (2), and downsloping (3) values.
12. Ca: Number of major vessels (0–3) colored by fluoroscopy.
13. Thal: This feature indicates the heart status with 3 values of normal (3), fixed defect (6), and reversible defect (7).
14. Num: It represents the diagnosis of heart disease with values of normal (0) and heart disease (1).

Proposed Clinical Decision Support System

Due to the asymptomatic nature of cardiovascular disease, its early diagnosis is crucial for saving patients' lives [22]. Accordingly, an effort was made to identify a pattern that can help identify individuals at a high risk of cardiovascular disease. This pattern was based on analyzing the characteristics found in the dataset of patient records.

There are multiple approaches for executing data mining projects, and one particularly effective method used in our research is the Crisp-DM (Crisp data mining) methodology [20]. We employed this methodology for cardiovascular disease prediction due to its ability to enhance the success rate of data mining projects. Crisp-DM allows the development and implementation of a robust data mining model applicable in real-world

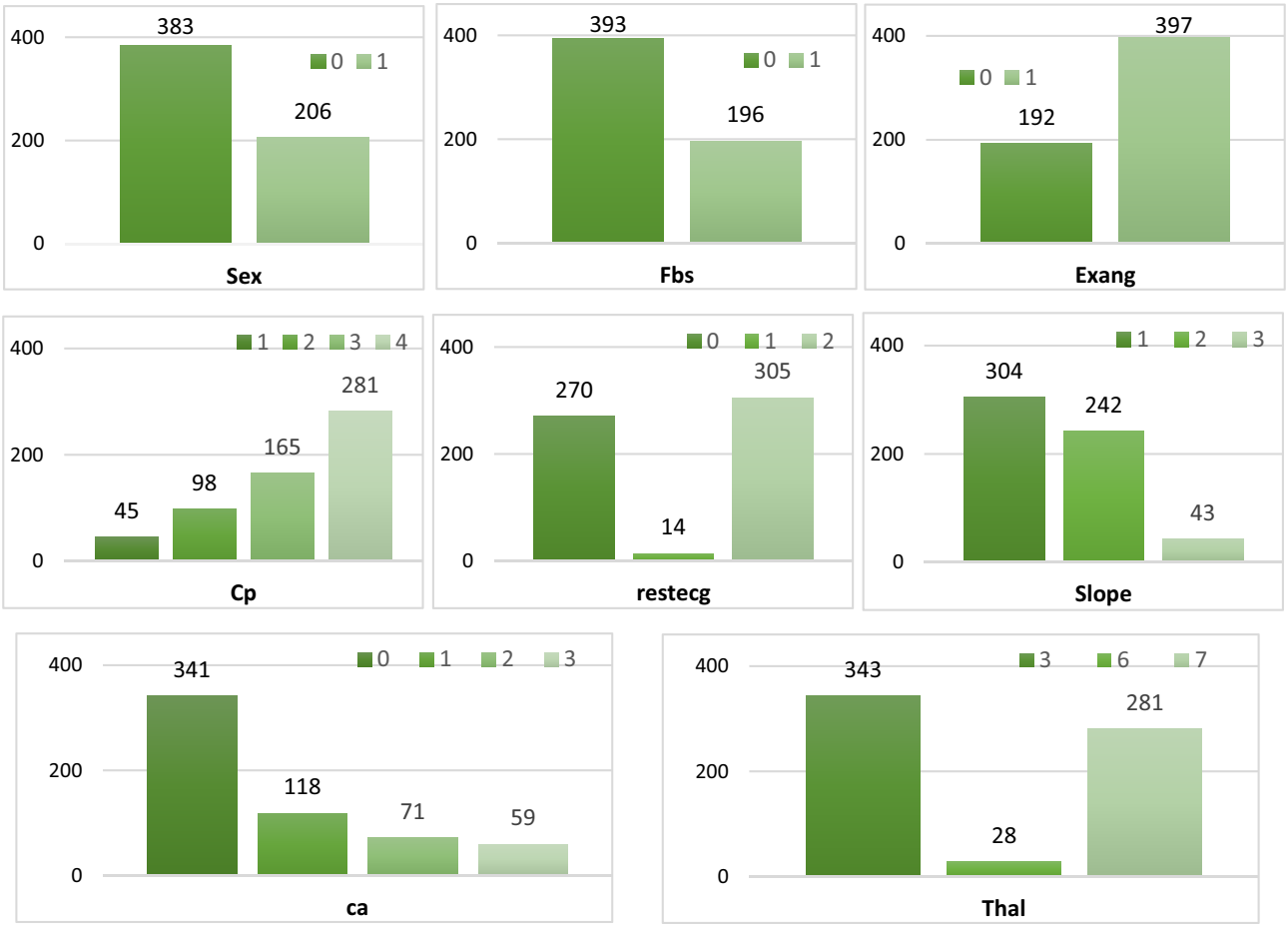


Fig. 3 Histogram of nominal features distribution in the Noor dataset

scenarios, enabling informed decision-making. Following the identification of targets, the methodology encompasses the following five phases. These phases are illustrated in Fig. 4, and each phase will be explained in more detail in the subsequent sections.

1. Data collection: The first step involves collecting the required data in accordance with the defined objectives.
2. Data preprocessing: To create an effective model, it is necessary to preprocess the collected data and extract relevant features.
3. Model development: This phase focuses on building a model that reflects the knowledge gained from the data.
4. System evaluation: The performance of the model is assessed and analyzed. If the model's accuracy falls short of expectations, alternative models are explored.
5. Development: If the performance of the generated model meets the desired standards, it can be deployed in a real-world setting.

Data Preprocessing

Data preprocessing is a crucial step in data mining which aims to prepare data for the important stage of the learning model. Its purpose is to decrease the number of attributes to enhance data quality and facilitate understanding of the rules generated by the models. However, it is important to note that only those features without a direct impact on the target attribute can be omitted. Initially, the gathered data were categorized into two groups: target variables and predictor variables, to identify the relevant attributes for model creation. In our study, the target feature was the heart disease diagnosis, while the remaining attributes were selected as predictors.

In the Cleveland dataset, there were six records with missing values. These records were removed from the dataset, decreasing the total number of records from 303 to 297. The target attribute, which indicated the absence or presence of heart disease, was originally represented by multiclass values (0 for absence and 1, 2, 3, and 4 for presence). It was

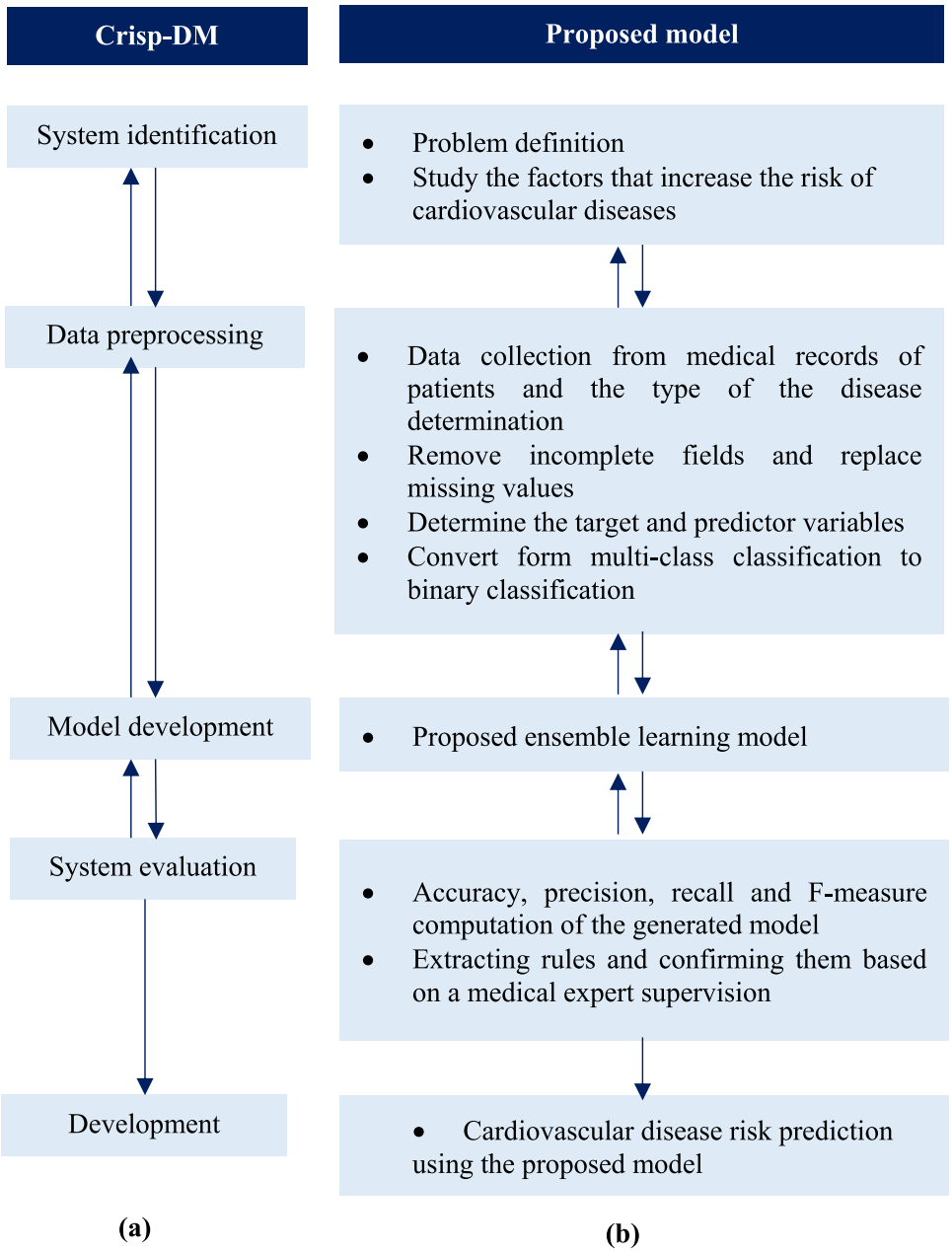


Fig. 4 Diagram of data mining model: Crisp-DM model (**a**) and proposed model (**b**)

transformed into binary values where 0 refers to the absence of heart disease and 1 refers to the presence of heart disease. During the preprocessing task, all diagnosis values ranging from 2 to 4 were converted to 1. As a result, the dataset only contained the values 0 and 1 for the diagnosis attribute, where 0 represents the absence of heart disease and 1 represents the presence of heart disease. Thereafter, the distribution of the 297 records for the "*Num*" attribute showed 160 records with a value of "0" (absence of heart disease) and 137 records with a value of "1" (presence of heart disease).

In the Noor dataset, there were 11 records that had missing values. These records were removed from the dataset, reducing the total number of records from 600 to 589. While the Cleveland dataset was transformed from multi-class classification to binary classification, the Noor dataset set initially had two labels for the target class. As a result, the dataset only contained the values 0 and 1 for the diagnosis attribute, where 0 represents the absence of heart disease and 1 represents the presence of heart disease. After removing missing values, the distribution of the 589 records for the "*Num*" attribute showed 332 records with a value of "0" (absence of heart disease) and 257 records with a value of "1" (presence of heart disease).

Table 2 Definition of voting algorithm rules

Rule	Definition	Formula
Majority	It is the most popular voting algorithm rule where the label that obtains more than 50% of the votes is selected	$\mu_j(x) = \sum_{t=1}^{T} d_{t,j}(x)$
Weighted majority	It is identical to the majority vote but each classifier effect is tuned using a separate weight	$\mu_j(x) = \sum_{t=1}^{T} w_t, d_{t,j}(x)$
Average of probabilities	It refers to the averaged probability of selecting each class as the final class by all classifiers	$\mu_j(x) = \frac{1}{T}\sum_{t=1}^{T} d_{t,j}(x)$
Product of probabilities	The probability of selecting a class as the target class is multiplied by all classifiers. The maximum product is then considered the final result	$\mu_j(x) = \prod_{t=1}^{T} d_{t,j}(x)$
Minimum probabilities	The minimum probability of each class is utilized as the target class	$\mu_j(x) = min_{t=1,\ldots,T}\{d_{t,j}(x)\}$
Maximum probabilities	The maximum probability of each class is utilized as the target class	$\mu_j(x) = max_{t=1,\ldots,T}\{d_{t,j}(x)\}$

Ensemble-Based Learning Model

There are various data mining algorithms that can be used for modeling, such as the Bayes' functions, meta, lazy, tree, and rule families. In this paper, we employed effective data mining algorithms to present a predictive model. To choose the best classifiers, we assessed the models on a development set, which is a common practice in data mining. Approximately 10% of the data were randomly selected as our development set. Thereafter, several classification algorithms, including Naïve Bayes and Bayesian network (Bayes family), support vector machine, multi-layer perceptron, and logistic regression (functions family), K-Star, IBK, and KNN (lazy family), decision table (rule family), and decision stump, J-48, and random tree (tree family) were experimented. These algorithms were chosen from a pool of over 40 algorithms due to their outstanding performance. Notably, we decided to include at least one algorithm from each family of classifiers to determine the optimal method within each family. All experiments were conducted using Python programming language based on *scikit-learn* tools. For detailed information about the mentioned algorithms, please refer to Han et al. [23], as their detailed explanations go beyond the scope of this paper.

While each classification algorithm was evaluated, the plan of combining these individual classifiers seemed beneficial. To accomplish this, the ensemble learning method was employed. Ensemble learning classifiers merge individual classifiers to classify new instances while the diversity and accuracy of classifiers are the essential requirements for combining different methods. Diverse methods produce varied outcomes when applied to new inputs, allowing the outputs to be combined to create improved classifiers. There are numerous approaches to construct ensembles with diverse classifiers. These approaches can be categorized into four levels such as classifier level, combination level, feature level, and data level [24, 25]. In the combination level, the focus is on developing different combiners while the classifier level leverages diverse base classifiers with distinct behavior. At the feature level, different subsets of features are employed, and dissimilar subsets of data are used at the data level [26, 27].

The combination level, which was used in our experiments, primarily concentrates on techniques for merging multiple base classifiers. At this stage, ensemble classifiers like bagging, boosting, and stacking were employed. The voting algorithm is a widely used method in ensembles [28]. There exist diverse voting algorithms with distinct rules for combining the classifiers. In a dataset, for every instance, the base classifiers assign probabilities to each class, determining the final class for that particular instance. These probabilities, which can be utilized within the voting algorithm's combination rule, are defined as $d_{t,j} \in [0,1] | t = 1, \ldots, T; j = 1, \ldots, C$ where T represents the total count of classifiers and C represents the number of different classes. $d_{t,j}$ refers to the likelihood or probability that classifier t selects class j as its outcome. Next, the voting algorithm calculates the combination rules based on the following Eq. (1). Notably, each $\mu_j(x)$ is calculated in each rule. Rule definitions are also provided in Table 2.

$$h_{final}(x) = arg_j max \mu_j(x) \tag{1}$$

The choice of optimal rules for the voting algorithm relies on the characteristics of the dataset. It is necessary to thoroughly analyze all rules in order to determine the most favorable outcome for the voting algorithm. Based on the empirical results, voting, boosting, and bagging methods exhibited better performance than individual classifiers.

Proposed Genetic-Based Ensemble Model

To improve the effectiveness of the voting algorithm, a weighted majority vote approach was employed. While a simple majority vote algorithm is generally efficient in combining diverse classifiers, it is important to acknowledge that not all classifiers have an equal impact on the classification task. To increase the outcomes of the weighted majority

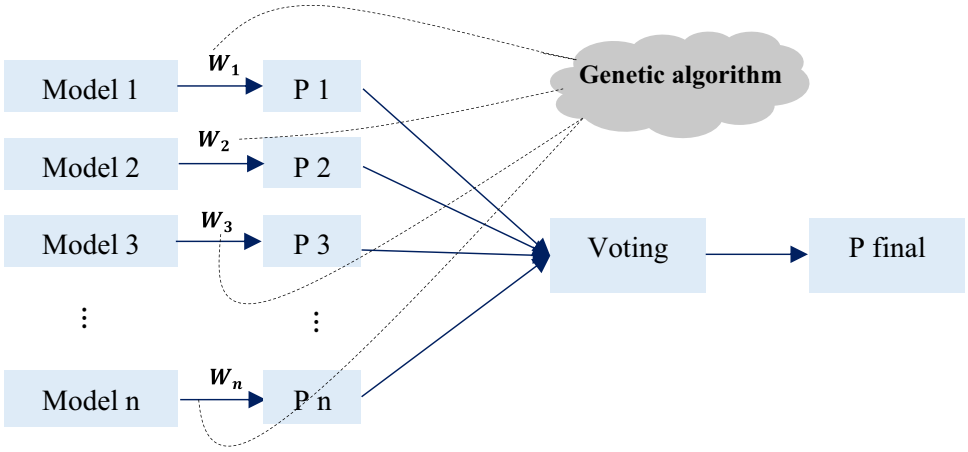


Fig. 5 Schematic structure of the proposed genetic-based ensemble learning model

vote classifier, it is crucial to identify the best weight vector. Accordingly, we decided to employ genetic algorithms [29] as an optimal solution for determining the most favorable weight vector.

Genetic algorithm is on the basis of Holland's evaluation theory [29] and has been utilized in extensive applications of various tasks and problems, particularly those that require optimizing multiple parameters. In the realm of machine learning and classification tasks, GA has been utilized for various tasks, including optimal features and classifier selection. Particularly, a weighted majority equation with T classifiers, where a vector of weight coefficients of size T serves as a chromosome in GA's population instances, was used in our proposed model. These population instances begin with randomly initialized weight vectors of varying values. During each generation, every instance is assessed using a fitness function, and the resultant outputs contribute to creating the next generation. The fittest chromosomes are kept, while others are removed. New instances predominantly arise from the best chromosomes, leading to the creation of the subsequent generation. The fitness function can take the form of a direct strategy rule, such as the output obtained from performing each weight vector. Ultimately, this algorithm yields the optimal weights for combining classifiers, which are then employed in the weighted majority vote classifier. The schematic structure is depicted in Fig. 5.

Chromosomes generated by the GA based on the proposed model's weighted majority vote algorithm are illustrated in Fig. 6. The weights assigned to each classifier range between 0 and 1. The genetic algorithm utilizes the provided weight vector as input. The fitness function evaluates the accuracy achieved by combining the classifiers utilizing the given weights on the development set. The population consists of 300 individuals initially generated with random weights assigned to each one. In every generation, the fitness function is applied to each member of the population, and the population is then sorted. The top 10% of the population is kept, and the subsequent 50% are served as parents for the next generation. New offspring is then generated by uniformly selecting weights from each parent (shown in the crossover row of Fig. 6).

To simulate an evolutionary process, a mutation step is performed. A random value ranging from − 1 to 1 is added to 0.05% of the population, considering the weight range. The optimal individual is chosen as the final vector for the weighted majority vote algorithm after 200 generations. In conclusion, the highest-performing algorithm for rank prediction is chosen and implemented in the web-based solution. When significant changes occur in the dataset, the whole process is repeated, resulting in the creation of new prediction models. Given the distinct characteristics and categories of different classifiers, it is anticipated that the weighted voting algorithm will yield improved results by effectively combining all classifiers with varying degrees of significance.

Fig. 6 Schematic structure of genetic algorithm including crossover and mutation steps. Each chromosome (each column) depicts the weight that is assigned to each classifier

	cw_1	cw_2	cw_3	cw_4	cw_5	cw_6	cw_7	cw_8	cw_9
Parent 1	0.925	0.425	0.665	0.157	0.229	0.328	0.810	0.896	0.728
Parent 2	0.543	0.803	0.108	0.854	0.651	0.328	0.605	0.653	0.542
Crossover	0.543	0.803	0.665	0.157	0.651	0.328	0.810	0.896	0.728
Mutation	0.543	0.956	0.665	0.157	0.651	0.528	0.810	0.896	0.728

Table 3 Summary of used hyperparameters

Methods		Hyperparameters
Individual classifier	Naïve Bayes	(alpha = 0.1, fit_prior = False, class_prior = [0.3, 0.7])
	Bayesian network	(estimator = MaximumLikelihoodEstimator, prior_type = 'BDeu', equivalent_sample_size = 10)
	MLP	hidden_layer_sizes = (100, 50), activation = 'relu', solver = 'adam', learning_rate = 'constant', alpha = 0.0001, batch_size = 'auto', max_iter = 200, random_state = 42)
	SVM	(C = 1.0, kernel = 'rbf', gamma = 'scale', degree = 3, coef0 = 0.0)
	Logistic regression	(penalty = 'l2', C = 1.0, solver = 'liblinear')
	K-star	(options = ["-E", "ModifiedEuclidean"], D = 'Z', B = ' Mean')
	IBK	(K = 3, distanceWeighting = ' Inverse '
	KNN	(n_neighbors = 5, weights = 'distance', algorithm = 'auto', metric = 'euclidean')
	Decision table	(criterion = 'gini', max_depth = None, min_samples_split = 2, min_samples_leaf = 1)
	Decision stump	(criterion = 'gini', splitter = 'best', min_samples_split = 2, min_samples_leaf = 1, max_features = None, random_state = 42)
	J-48	(options = ["-C", "0.25", "-M", "2"])
	Random tree	(n_estimators = 100, max_depth = None, min_samples_split = 2, min_samples_leaf = 1, random_state = 42)
Ensemble-based methods	Adaboost	(n_estimators = 50, learning_rate = 1.0, algorithm = 'SAMME.R', random_state = 42)
	Bagging	(n_estimators = 10, max_samples = 1.0, max_features = 1.0, bootstrap = True, random_state = 42)
	Logic boost	(n_estimators = 100, learning_rate = 0.1, max_depth = 3, subsample = 1.0, random_state = 42)
	XGboost	(n_estimators = 100, learning_rate = 0.1, max_depth = 3, subsample = 1.0, random_state = 42)
	Voting	(voting = 'soft', weights = [1, 1])
Proposed model		("population_size" = 300, "generations" = 200, "crossover_rate" = 0.1, "mutation_rate" = 0.05, "voting_weights" = [0.5, 0.3, 0.2])

Experiment and Results

Evaluation Metrics

Knowledge generated in the previous step must be carefully examined and interpreted. The objective of knowledge evaluation is to specify its accuracy and suitability for practical applications. Various methods are employed to assess the generated knowledge, which is tied to the used learning models. In order to handle overfitting, the tenfold cross-validation technique, a widely accepted method for assessing classification algorithms, was utilized in our evaluations which not only helps to assess model performance on different subsets of data but also reveals how well the model is generalizable to unseen data. We employed four standard metrics, namely accuracy, precision, recall, and *F*-measure, to assess the effectiveness of the proposed model. Additionally, we incorporated the AUC (area under the ROC curve) metric that is commonly used in medical data mining tasks. These metrics can be computed based on the following Eqs. (2–5). The subsequent section illustrates the results of each separate classifier as well as the combined ensemble classifiers using the following equations.

$$accuracy = \frac{number\_of\_correctly\_predicted\_sample}{total\_number\_of\_samples} \quad (2)$$

$$precision = \frac{number\_of\_correctly\_predicted\_samples}{number\_of\_predicted\_samples} \quad (3)$$

$$recall = \frac{number\_of\_correctly\_predicted\_samples}{number\_of\_correct\_samples} \quad (4)$$

$$F - measure = \frac{2 \times P \times R}{P + R} \quad (5)$$

Hyperparameters

Hyperparameters are configuration settings that are external to a model that influence the behavior of the algorithm and can significantly impact the model's performance and generalization ability. Therefore, tuning hyperparameters is vital as they directly influence the model's performance and predictive capabilities. Hyperparameters act as tuning knobs that control the behavior and complexity of the model, impacting its ability to capture underlying patterns in the data. While we trained several classification algorithms in our study, we tried our best to set hyperparameters properly to fine-tune each algorithm for optimal performance and ensure that it can effectively capture patterns within the

Table 4 Precision, recall, F1, and AUC measures on both Cleveland and Noor datasets

Methods		**Cleveland dataset**				**Noor dataset**			
		Precision	Recall	F1	AUC	Precision	Recall	F1	AUC
Individual classifier	Naïve Bayes	84.05	82.29	82.47	88.12	86.05	84.34	84.73	90.54
	Bayesian network	82.98	81.31	81.43	85.34	84.63	83.14	83.45	89.18
	MLP	83.12	79.92	81.56	86.28	85.73	81.63	83.14	89.64
	SVM	82.85	81.20	80.73	86.48	84.63	83.14	82.18	88.34
	Logistic regression	83.23	82.96	82.78	87.14	85.17	84.18	84.63	89.41
	K-star	70.28	71.53	70.63	75.13	73.89	74.64	74.34	77.56
	IBK	76.23	77.51	76.34	80.63	79.53	79.51	78.68	83.11
	KNN	77.61	75.84	78.64	81.24	80.12	78.63	80.14	84.97
	Decision table	78.36	77.61	78.96	81.64	80.63	79.85	81.54	84.34
	Decision stump	71.29	72.64	72.88	75.38	73.43	74.65	74.91	78.17
	J-48	84.63	84.18	81.48	86.73	88.14	86.92	83.95	91.15
	Random tree	78.57	77.55	78.45	82.43	81.07	80.85	80.71	84.63
Ensemble-based methods	Adaboost	85.16	85.24	85.04	89.57	87.32	87.64	87.31	91.64
	Bagging	85.97	86.05	86.14	89.87	87.21	88.34	87.73	92.14
	Logic boost	86.33	86.14	86.71	90.01	88.32	88.21	88.46	92.11
	XGboost	86.14	86.37	86.14	90.91	88.31	88.43	88.34	92.84
	Voting	87.23	87.43	87.61	89.43	89.14	89.03	89.71	93.05
Proposed model		88.05	87.63	87.91	91.34	90.12	89.32	89.73	94.14

data. The summary of used hyperparameters is provided in Table 3.

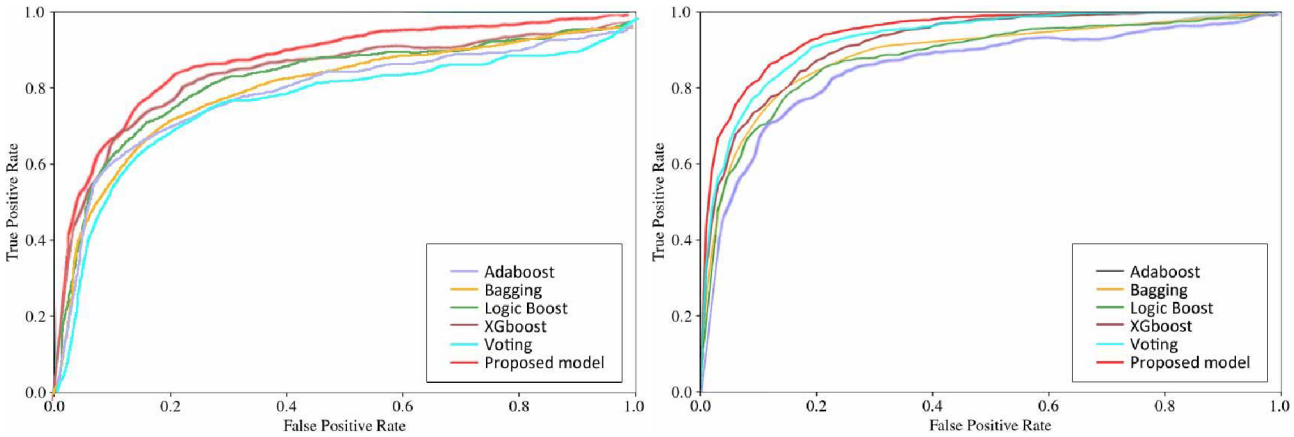
Performance Evaluation

The goal of this paper is to introduce a model that utilizes data mining algorithms to predict the risk of cardiovascular diseases. Accordingly, various data mining algorithms besides ensemble-based methods were implemented on both datasets. The results of empirical experiments are presented in Table 4. Based on the result of the experiments, it can be stated that:

- Leveraging ensemble-based methods, including AdaBoost, Voting, Logit Boost, and Bagging led to greater results compared to single-classification methods. Among all ensemble-based methods, *voting* presented the highest precision.
- The conclusion can be drawn from the last line of Table 4, which demonstrates the superiority of our proposed model. It highlights the efficiency of our ensemble-based learning model compared to traditional ensemble learning methods. Essentially, our model has higher precision and can be served as a benchmark for future research. The study's findings present a comprehensive model (with a precision of 88.05% and 90.12% on Cleveland and Noor datasets, respectively) for cardiovascular disease diagnosis using the previously described features. These results emphasize that intelligently weighting individual classifiers is an efficient approach to combine classifiers in ensemble-based methods.
- For a better illustration of the superiority of our proposed model compared to traditional ensemble learning methods, their ROC curves on both datasets are illustrated in Fig. 7 to provide a visual representation of how well various models can distinguish between the two classes and offer insights into their discriminatory ability. As can be seen, the proposed model has a higher AUC which indicates that it has better discriminatory power and is able to distinguish between the classes more effectively.

Subset Selection

Feature selection is the initial step in any data mining task. Therefore, conducting an essential experiment to evaluate the impact of individual features is another crucial aspect of addressing the given problem. To this end, the correlation between attributes was computed and ranked to identify the most important features. Table 5 and Fig. 8 show the effect of different features on both datasets. As illustrated, *sex*, *ca*, and *cp* are the most important features on both datasets while *age* is the least important one on both.



(a) Cleveland dataset (b) Noor dataset

Fig. 7 ROC curves of different ensemble learning methods in comparison to the proposed model

Generated Rules

As previously mentioned, the model created with the J-48 algorithm exhibited a high level of precision. While our study focuses on predicting the risk of cardiovascular disease using patient medical records to assist specialists, certain rules were extracted from the model. These rules, which are presented in Table 6, can be utilized by specialists to make advanced predictions about the diagnosis of cardiovascular disease.

Table 5 Different attribute effects on heart disease on Cleveland and Noor datasets

Cleveland dataset		Noor dataset	
Feature	Effect	Feature	Effect
Sex	0.64218	Sex	0.71245
Ca	0.58914	Ca	0.65145
Cp	0.50483	Cp	0.54325
Oldpeak	0.47015	Slope	0.35841
Thal	0.30241	Thal	0.30478
Slope	0.21047	Oldpeak	0.20143
Exang	0.11024	Fbs	0.15873
Fbs	0.09841	Exang	0.09347
Restecg	0.08312	Restecg	0.08573
Thalach	0.03186	Thalach	0.04314
Chol	0.02115	Trestbps	0.02141
Trestbps	0.01015	Chol	0.01810
Age	0.00303	Age	0.00211

Running Time

The running time of our proposed method can be considered one of its intriguing aspects. The test step, which is done online, took approximately 0.038 s for every patient which is favorable and can be considered a real-time operation.

Discussion

Data mining represents a potent and innovative technology for uncovering concealed predictive and actionable information from extensive databases, enabling profound and original insights. Employing advanced data mining methods to extract valuable information has been regarded as a proactive strategy to enhance the quality and precision of healthcare services, while simultaneously reducing healthcare costs and diagnosis time [30].

Noteworthy, accurate diagnosis of cardiovascular disease is crucial for planning appropriate care. For accurate cardiovascular disease diagnosis, clinicians need to consider the patients' health history and the results of their recent clinical tests. The task of making these decisions accurately and efficiently is quite challenging for healthcare practitioners, as even a slight oversight can put the patient's life at risk [31]. However, using data mining can help specialists in making correct decisions. To this end, data mining techniques were used in this paper to develop an appropriate model for cardiovascular disease prediction. Accordingly, locally collected dataset besides the Cleveland dataset was used in our experiments to determine the most important features and the most suitable data mining methods.

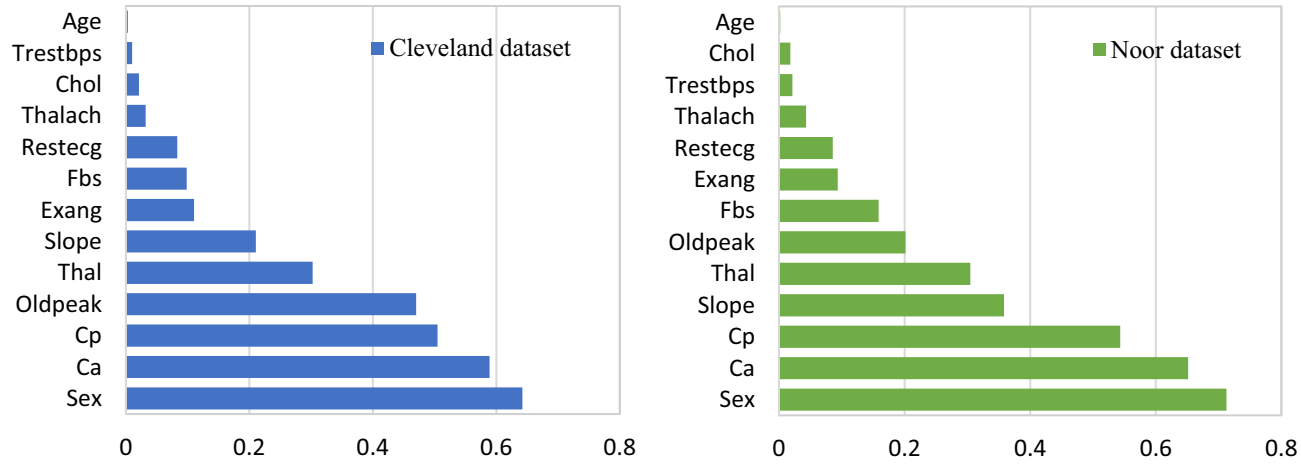


Fig. 8 Histograms of various attribute effects on heart disease on Cleveland and Noor datasets

To this end, Crisp-DM methodology was used to increase the success rate of data mining methods. Thereafter, various data mining techniques were first implemented on both datasets and then the voting algorithm, a representative of ensemble learning, was used to combine individual classification methods to classify new instances. A weighted majority vote based on the genetic algorithm was also utilized to increase the voting algorithm's performance. In summary, utilizing suitable retrospective medical datasets, the proposed data mining model established a decision support system that predicted the presence or absence of heart disease during the treatment phase. Based on the generated models, the following features, namely *sex*, *ca*, *cp*, *oldpeak*, *thal*, *slope*, *exang*, *fbs*, and *restecg*, were the most effective features in predicting cardiovascular disease. Among the implemented methods, the proposed ensemble-based model also had the highest classification performance. However, the process does not end with generating the model and the generated knowledge must be organized to enhance its usefulness. Noteworthy, to verify the validity of the generated rules, they were presented to the cardiologist, and their correctness was confirmed.

To future analyze the superiority of the proposed model, a benchmark comparison is required. Benchmarking serves as a valuable tool for evaluating the performance of a particular model in comparison to others. Accordingly, the

Table 6 Sample rules generated by the J-48 model

Row	Class	Risk of illness	Rules
1	Absence of heart disease	0.99	If the *sex* is female, *fbs* is false, *restecg* is normal, *exang* is false, and *thal* is normal, the person is healthy
2		0.98	If the *sex* is female, *cp* is non-anginal pain, and *thal* is normal, the person is healthy
3		0.93	If the *sex* is male, *thal* is normal, and *cp* is zero, the person is healthy
4		0.92	If *sex* is male, oldpeak = (-inf, 0.56), *cp* is zero, and *thal* is normal, the person is healthy
5		0.91	If the *sex* is female, *cp* is non-anginal, and *thalach* is (149.6,175.8), the person is healthy
6		0.90	If *tresrbps* less or = (115.2,136.4), *exang* is false, *cp* is zero, and *thal* is normal, the person is healthy
7	Presence of heart disease	0.99	If the *sex* is male, *cp* is asymptomatic, and *ca* is two, the person is sick
8		0.098	If *thalach* = (123.4,149.6), *exang* is true, and *thal* is reversible, the person is sick
9		0.96	If *cp* is asymptomatic, *slope* is flat, and *thal* is reversible the person is sick
10		0.94	If *cp* is asymptomatic, *exang* is true, and *thal* is reversible, the person is sick
11		0.92	If *sex* is male, *cp* is asymptomatic, *fbs* is false, *exang* is true, and *thal* is reversible, the person is sick
12		0.91	If *cp* is asymptomatic, *exang* is true, and *slope* is flat, the person is sick
		0.90	If *sex* is female, *cp* is asymptomatic, *restecg* is high, *slope* is flat, and *thal* is reversible, the person is sick

Table 7 Comparison of the proposed method with state of the arts on the Cleveland dataset

Source	Method	Accuracy (%)
Proposed model	Ensemble model based on GA	88.43%
Ahmad et al. (2023) [30]	SVM	87.91%
Akkaya et al. (2022) [31]	KNN	85.60%
Tougui et al. (2020) [34]	Random forest	87.64%
Shafenoor et al. (2019) [35]	Vote with Naïve Bayes and logistic regression	87.41%
Subanya and Rajalaxmi (2014) [36]	SVM	86.76%
Mokeddem et al. (2013) [37]	GA + Naive Bayes	85.50%
Khanna et al. (2015) [38]	Logistic regression	84.80%
Mokeddem et al. (2013) [37]	GA + SVM	83.82%
Kumar et al. (2018) [39]	Decision tree C4.5	83.40%
Acharya (2017) [40]	KNN	82%

proposed model was compared to the state of the arts to confirm whether it achieved a satisfactory level of accuracy when compared to the accuracy attained by previous studies conducted on the Cleveland dataset. In this regard, Ahmad et al. [32] trained six machine learning algorithms including logistic regression, K-nearest neighbor, SVM, decision tree, random forest classifier, and extreme gradient boosting on two heart disease datasets. Based on the result of their experiments, SVM obtained the highest accuracy of 87.91% on the Cleveland dataset. Akkaya et al. [33] analyzed eight different machine learning classification methods on the Cleveland dataset and concluded that KNN with the accuracy of 85.6% had the best performance. Tougui et al. [34] also implemented various data mining methods. Based on the result of their experiments, random forest obtained the highest classification accuracy of 87.64%. Moreover, Shafenoor et al. [35] investigated the efficiency of the data mining techniques to specify the important features besides the classification of the presence or absence of heart diseases. They concluded that voting with Naïve Bayes and logistic regression has the highest classification accuracy of 87.41%. Following a similar line of research, Subanya and Rajalaxmi [36] utilized SVM classification method besides the Swarm intelligence-based artificial bee colony (ABC) algorithm to find the best features and obtained an accuracy of 86.76%. Moreover, Mokeddem et al. [37] employed the genetic algorithm along with Naïve Bayes and SVM to perform classification and achieved an accuracy of 85.50% and 83.82%, respectively. Khanna et al. [38] conducted a comparative study of classification techniques (SVM, logistic regression, and neural networks) to predict the prevalence of heart disease and concluded that logistic regression with a classification accuracy of 84.80% had the best performance. Moreover, Kumar et al. [39] implemented eight various data mining methods to predict heart disease and concluded that decision tree C4.5 with an accuracy of 83.40% had the best performance. Acharya [40] also investigated the efficiency of various data mining techniques to predict the presence of heart disease and concluded that KNN is the best algorithm with a classification accuracy of 82%. As can be seen, our proposed model with a classification accuracy of 88.43% has superior performance compared to state of the arts which clearly demonstrates its potential in predicting the presence or absence of heart disease based on clinical features. The comparison results are provided in Table 7.

It is worth mentioning that although the proposed model developed using the mentioned data presented superior performance, it may not generalize well to different settings or populations. Variations in patient demographics, healthcare practices, and treatment protocols can affect the performance of predictive models when applied in different contexts. Furthermore, the data used for developing the proposed model may not fully represent the population at large. Considering the fact that disease patterns and risk factors may evolve over time due to various factors such as lifestyle changes, medical advancements, and population demographics, the proposed model developed using the mentioned historical data may struggle to adapt to these changing patterns which may result in decreased prediction accuracy.

Conclusion

Cardiovascular diseases are the leading cause of death worldwide. Therefore, its early detection is of paramount importance in healthcare. Data mining plays a significant role in this field by identifying risk factors, enabling predictive analytics, supporting decision-making, and facilitating knowledge discovery, thereby contributing to more proactive and personalized approaches to heart disease management. Accordingly, an ensemble-based model for precise forecasting of cardiovascular disease and pinpointing significant factors that have the highest influence was proposed in this paper. Different data mining methods along with four

ensemble learning models were applied to both sets of data. To this end, a new approach for merging individual classifiers in ensemble learning, where weights were given to each classifier (using a genetic algorithm), was created. The effectiveness of each attribute for prediction was also examined to confirm the reliability of the results.

To prove the efficiency of the proposed model, the Cleveland dataset besides the locally collected dataset, called the Noor dataset, was used in our experiments. To conduct a meaningful comparison, the datasets were initially subjected to the same data mining methods. Based on the results of experiments on both datasets, the proposed model presented superior performance compared to both individual and ensemble-based classifiers. The findings of this study put forward an accurate model that can be employed to predict the risk of cardiovascular disease which can be crucial for effectively managing and utilizing resources. Moreover, it serves as a valuable tool for cardiologists and physicians in classifying new patients besides estimating the needed human resources like doctors, technicians, nurses, and essential medical equipment.

There are numerous possibilities for improving this research and overcoming the limitations of this study. One approach is to expand the scope by conducting the same experiment on larger real-world datasets. Further investigation can explore different combinations of data mining methods for predicting cardiovascular disease. Additionally, applying new feature selection methods can provide a wider understanding of the important features, thereby enhancing prediction accuracy. Employing the proposed method in other domains is worth exploring and can be considered a possible future work.

Author Contributions Mojdeh Nazari and Reza Rabeie conceived the presented idea. Mojdeh Nazari developed the theory and performed the computations. Hassan Emami conceived the study and was in charge of overall direction and planning. Azamossadat Hosseini and Shahabedin Rahmatizadeh verified the analytical methods and obtained results. All authors discussed the results and contributed to the final manuscript.

Funding This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of Data and Materials Two various datasets were used in our experiments. The first one is the Cleveland dataset collected from the UCI machine learning repository which is available online. The second dataset is the "Noor dataset" which is freely available for academic purposes upon request.

Declarations

Ethical Approval There are no human or animal subjects in this study. It is not applicable.

Competing Interests The authors declare no competing interests.

References

1. Bhatt CM, et al. Effective heart disease prediction using machine learning techniques. Algorithms. 2023;16(2):88.
2. Ramesh T, et al. Predictive analysis of heart diseases with machine learning approaches. Malays J Comput Sci. 2022;132–48.
3. Nagavelli U, Samanta D, Chakraborty P. Machine learning technology-based heart disease detection models. J Healthc Eng. 2022;2022.
4. Al-Jammali K. Prediction of heart diseases using data mining algorithms. Informatica. 2023;47(5).
5. Tsao CW, et al. Heart disease and stroke statistics—2023 update: a report from the American Heart Association. Circulation. 2023;147(8):e93–621.
6. Bakar WAWA, et al. A review: heart disease prediction in machine learning & deep learning. in 2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA). 2023. IEEE.
7. Mohades Deilami F, Sadr H, Tarkhan M. Contextualized multi-dimensional personality recognition using combination of deep neural network and ensemble learning. Neural Process Lett. 2022;54(5):3811–28.
8. Ogunpola A, et al. Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics. 2024;14(2):144.
9. Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl. 2018;29:685–93.
10. Bhatt A, et al. Data mining approach to predict and analyze the cardiovascular disease. in Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications: FICTA 2016, Volume 1. 2017. Springer.
11. Kodati S, Vivekanandam R. Analysis of heart disease using in data mining tools Orange and Weka. Glob J Comput Sci Technol C Softw Data Eng. 2018;18(1):16–22.
12. Garate Escamilla AK, Hajjam El Hassani A, Andres E. Acomparison of machine learning techniques to predict the risk of heart failure. Machine Learning Paradigms: Applications of Learning and Analytics in Intelligent Systems. 2019;9–26.
13. Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Inf Med Unlocked. 2019;16:100203.
14. Garavand A, et al. The attributes of hospital-based coronary artery diseases registries with a focus on key registry processes: a systematic review. Health Inf Manag J. 2022;51(2):63–78.
15. Alotaibi N, Alzahrani M. Comparative analysis of machine learning algorithms and data mining techniques for predicting the existence of heart disease. Int J Adv Comput Sci Appl. 2022;13(7).
16. Ayatollahi H, Gholamhosseini L, Salehi M. Predicting coronary artery disease: a comparison between two data mining algorithms. BMC Public Health. 2019;19(1):1–9.
17. Alizadehsani R, et al. Machine learning-based coronary artery disease diagnosis: a comprehensive review. Comput Biol Med. 2019;111:103346.
18. Garavand A, et al. Designing the coronary artery disease registry with data management processes approach: a comparative systematic review in selected registries. Int Cardiovasc Res J. 2020;14(1).
19. Zahmatkesh Zakariaee A, Sadr H, Yamaghani MR. A new hybrid method to detect risk of gastric cancer using machine learning techniques. J AI Data Min. 2023;11(4):505–15.
20. Wirth R, Hipp J. CRISP-DM: Towards a standard process model for data mining. in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. 2000. Manchester.
21. Dua D, Graff C. UCI machine learning repository. School of Information and Computer Science, University of California, Irvine,

CA. School of Information and Computer Science University of California Irvine CA, 2019.
22. Ozcan M, Peker S. A classification and regression tree algorithm for heart disease modeling and prediction. Healthcare Analytics. 2023;3:100130.
23. Han J, Kamber M, Pei J. Data mining concepts and techniques third edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University, 2012.
24. Sagi O, Rokach L. Ensemble learning: a survey. Wiley Interdiscip Rev: Data Min Knowl Discov. 2018;8(4):e1249.
25. Kuncheva LI. Combining pattern classifiers: methods and algorithms. 2014: John Wiley and Sons.
26. Dietterich TG. Ensemble methods in machine learning. in International workshop on multiple classifier systems. 2000. Springer.
27. Akbar S, et al. pAtbP-EnC: identifying anti-tubercular peptides using multi-feature representation and genetic algorithm based deep ensemble model. IEEE Access. 2023.
28. Kim H, et al. A weight-adjusted voting algorithm for ensembles of classifiers. J Korean Stat Soc. 2011;40(4):437–49.
29. Sampson JR. Adaptation in natural and artificial systems (John H. Holland). 1976, Society for Industrial and Applied Mathematics.
30. Ali MM, et al. Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison. Comput Biol Med. 2021;136:104672.
31. Ali F, et al. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Inform Fusion. 2020;63:208–22.
32. Ahamad GN, et al. Influence of optimal hyperparameters on the performance of machine learning algorithms for predicting heart disease. Processes. 2023;11(3):734.
33. Akkaya B, Sener E, Gursu C. A comparative study of heart disease prediction using machine learning techniques. In 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). 2022. IEEE.
34. Tougui I, Jilbab A, El Mhamdi J. Heart disease classification using data mining tools and machine learning techniques. Health Technol. 2020;10:1137–44.
35. Amin MS, Chiam YK, Varathan KD. Identification of significant features and data mining techniques in predicting heart disease. Telemat Inform. 2019;36:82–93.
36. Subanya B, Rajalaxmi R. Feature selection using artificial bee colony for cardiovascular disease classification. in 2014 International Conference on Electronics and Communication Systems (ICECS). 2014. IEEE.
37. Mokeddem S, Atmani B, Mokaddem M. Supervised feature selection for diagnosis of coronary artery disease based on genetic algorithm. arXiv preprint arXiv:1305.6046, 2013.
38. Khanna D, et al. Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease. Int J Mach Learn Comput. 2015;5(5):414.
39. Kumar MN, Koushik K, Deepak K. Prediction of heart diseases using data mining and machine learning algorithms and tools. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2018;3(3):887–98.
40. Acharya A. Comparative study of machine learning algorithms for heart disease prediction. 2017.